

Unit 4

Lesson 1

Parsing Data for Data Cleanse

LESSON OVERVIEW

In this lesson, you will learn how to parse data for data cleanse.



LESSON OBJECTIVES

After completing this lesson, you will be able to:

- Parse data for cleansing

Data Cleanse Parsing

The Data Cleanse transform identifies and parses names, titles, firm data, phone numbers, Social Security numbers, dates, and e-mail addresses. You can also assign gender code, add prenames, create personalized greetings, generate match standards, and convert input sources to a standard format. Additionally, you can parse multiple names from individual records so that records can be created for each individual.

There are five main steps that the Data Cleanse transforms takes while parsing operational data:

1. Word breaking – breaks the input line down into smaller, more usable pieces. Data Cleanse breaks the input line on white space, punctuation, and alphanumeric transitions.
2. Gathering – recombines words that belong together, such as words that look up together in the dictionary. Data Cleanse does not attempt to combine words that have been broken for a custom parser.
3. Tokenization – assigns specific meanings to each of the pieces. Data Cleanse looks up each individual input word in the dictionary. A list of tokens is created using the classifications associated with each word in the dictionary.
4. Rule matching – matches the token meanings against defined rules. Data Cleanse does not match the pattern of specific words against the rules; it matches pattern of the types, or classifications, of the words.
5. Action item assignment – outputs parsed data based upon matched rules.

The Business Need for Data Cleanse Transforms

As with address cleansing, one of the primary uses of data cleansing is to prepare names, titles, firm data, phone numbers, Social Security numbers, dates, and e-mail addresses for matching. Dictionaries are used with rule files to parse data to its smallest components and standardize it. This reduces variability and increases the possibility of a successful match.

In addition, data cleansing can generate match name standards, which provide:

- Alternative spellings for names (for example, Kathy and Cathy)
- Alternatives for ambiguous names (for example, Patrick and Patricia for Pat)